利用引文构建的主题模型研究进展*

■ 邹丽雪[1,2] 王丽[1,2] 刘细文[1,2]
[1] 中国科学院文献情报中心 北京 100190
[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘要：[目的/意义]概率主题模型算法在不断得到改进与扩展，本文对国内外已有的利用引文构建的主题模型进行研究，分析和对比不同模型的生成过程与算法，并探讨利用引文构建的主题模型在科技文本分析中的应用与可扩展的研究方向。[方法/过程]通过 Web of Science 数据库和 CNKI 数据库获取国内外利用引文构建主题模型的相关文献，经人工判读后筛选出具有代表性的文献，对这些文献中利用引文构建的主题模型，从建模思想、生成过程、参数估计与推断算法等方面进行对比与分析。[结果/结论]目前国内外利用引文构建的主题模型主要包括研究主题与引文分布的主题模型、研究被引与施引主题间关系的主题模型，以及基于引用内容的引用主题模型；主题模型中引入引文信息后，能够获得更完整的主题内容和特定主题下的重要文献，并可识别施引文献和被引文献之间主题间的关系及影响；已有的模型多集中在概率潜在语义分析(Probabilistic Latent Semantic Analysis，PLSA)和潜在狄利克雷分配(Latent Dirichlet Allocation，LDA)主题模型基础上进行扩展。未来可扩展研究引入引用内容的主题模型、模型的性能优化和评价方法、模型的应用研究等。

关键词：主题模型 引文 主题识别 引用内容
分类号：G202
DOI：10.13266/j.issn.0252-3116.2019.23.015

1 引言

信息化时代，以文本资源为典型的各种信息呈爆炸式增长，信息的不断积累导致文本的数据量日益庞大。其中，科技文献的数量呈指数倍增长，科技文献作为知识的主要载体，是知识发展过程的累积形态，蕴含着大量揭示学科发展演化的主题信息。从主题复杂多样且数据庞大的文本信息中挖掘出其蕴含的主题及主题演化信息，可以帮助科研人员以及决策人员识别学科领域研究的主题内容、快速了解与把握科技发展的脉络、跟踪科技领域主题的演化状态和知识流动的轨迹。

近年来在文本挖掘领域，概率主题模型[1]是机器学习和自然语言处理领域中用于在一系列文档中发现隐含主题的一种统计模型，可实现文本语义挖掘。作为一套新的能对文献进行语义抽取的算法，概率主题模型引入主题空间的概念，实现了文档在主题空间上的表示，而每一个主题又能够表示成为一个在词袋空间上的概率分布。与空间向量和语言模型不同的是，概率主题模型引入主题概念后，不仅能够实现文档的降维表示，同时能够抽取出文档集合上的隐含语义，为大规模数据集中的文档寻找一个相对短的描述[2]。通过对文本中深层的、隐含的语义信息进行挖掘，能够更好地从科研文献中抽取出更有价值的潜在主题分布，这种新的潜在的语义空间在文档和词之间填补了空白，提供了一种帮助科研人员在大量文本中识别主题的新方法，已成为非常活跃的研究领域。随着自然语言处理技术的发展，概率主题模型被广泛地应用于主题识别和主题演化等领域中。

概率主题模型有很多算法，经典的两种算法为概率潜在语义分析(Probabilistic Latent Semantic Analysis，PLSA)[3]和潜在狄利克雷分配(Latent Dirichlet Allocation，LDA)主题模型[1]。LDA 由于具有良好的数学基础和灵活的扩展性，得到了广泛地应用与扩展。但随

* 本文系中国科学院文献情报中心青年人才领域前沿项目“基于引用内容关联的多维主题演化研究”(项目编号：G1726)研究成果之一。
作者简介：邹丽雪(ORCID：0000-0002-2617-4151)，馆员，博士研究生；王丽(ORCID：0000-0002-9513-6159)，副研究馆员，博士研究生；刘细文(ORCID：0000-0003-0820-3622)，研究员，博士生导师，通讯作者，E-mail：liuxw@mail.las.ac.cn。
收稿日期：2019-01-28 修回日期：2019-06-25 本文起止页码：131-138 本文责任编辑：王传清

着研究的深入,学者们也指出了 LDA 模型存在的问题,比如 LDA 模型建立时假设文档之间可交换即认为文档之间没有先后顺序、主题之间可交换即认为主题之间没有层次关系和先后关系[4]。然而,大部分语料库特别是科技文献之间以很多方式相互关联而不是独立的,文档中主题的产生往往有先后顺序和层次关系,显然这些假设没有对主题的关联关系进行建模,在分析语料库时应考虑这样的信息。

随后研究人员对模型算法进行了改进与扩展,包括对主题演化过程引入时间因素进行建模[5],引入作者元数据构建作者主题模型[6]等。一篇研究论文除了是一个词袋,还包含了更多的结构信息,其中引文作为在科学文献中重要的继承元素,所包含的噪声信息会更少,且能说明一个文档对另一个文档的影响以及主题间的联系,因此,国内外学者将引文关联关系引入到主题模型中,对模型算法进行了改进与扩展。本文针对国内外已有的利用引文构建的主题模型进行深入的分析和对比,详细阐述这类模型的生成过程与算法,同时指出存在的问题,并探讨利用引文构建的主题模型在科技文本分析中的应用与今后可扩展的研究方向。

2 数据来源及研究方法

为了全面分析目前引入引文的主题模型的最新研究进展,本文选取 Web of Science 数据库核心合集数据库、中国知网数据库(CNKI)作为数据来源分别进行英文、中文文献检索。数据的获取过程为:①以主题为((“topic model * ” and (citation or citations or cited or citing or reference *)) or ((“bayesian model * ” or “probabilistic model * ”) near (citation or citations or cited or citing or reference *)) or ((model * near topic *) near (citation or citations or cited or citing)))进行英文文献检索,文献类型限定为论文(article)、会议论文(proceeding paper)、评论(review)和社论材料(editorial material),检索时间截至 2019 年 1 月 28 日,检索到英文文献 381 篇,其中论文 229 篇,会议论文 149 篇,从发文的主要国家看,美国学者发表 129 篇,中国学者发表 93 篇;②以主题为(模型 * (引用 + 引文) * (主题 + 概率 + 贝叶斯))进行中文文献检索,中文文献类型限定为期刊论文、会议论文、学位论文,检索时间截至 2019 年 1 月 28 日,检索到中文文献 155 篇;③对两个数据集中的文献分别进行人工逐一判读,筛选去掉内容相关度较小的文献,最终选取将引文引入主题模型中对模型进行改进的具有代表性的 26 篇文献进行分析。本文将分别从模型的建模思想、生成过程、模型参数估计与推断的算法、优势与不足等角度对已有的利用引文构建的主题模型进行详细对比与分析,并提出未来的可能发展趋势。

3 利用引文构建的主题模型进展分析

本文针对上述 26 篇代表性文献的研究内容,对其提出的基于引文构建的主题模型按研究角度划分为 3 个方向,包括研究主题与引文分布的主题模型、研究被引与施引主题间关系的主题模型,以及基于引用内容的引用主题模型,本文将进一步对这些利用引文构建的主题模型进行详细的分析和对比。

3.1 基于主题与引文分布的主题模型

该类主题模型引入了引文,研究了引文的主题分布,抽取引文文档与主题的分布,或将引文作为单词一样引入,从而获得主题与引文的分布。该类模型主要包括 PHITS、PLSA-PHITS、Mixed-membership model、cc-LDA、cp-LDA、CitationLDA + + 、Citation Author Topic Model(CAT)、Citation Topic Model(CT)、Citation-Content-LDA、Citation Network Topic Model(CNTM)等。

早期将引文与文本内容联合建模是对 PLSA 模型进行扩展,D. Cohn 等[7]在 PLSA 模型的基础上借鉴超链接引导的主题搜索算法(Hyperlink-Induced Topic Search,HITS)扩展得到了 PHITS 模型。该模型假设引用的生成过程类似于 PLSA,不同之处在于 PLSA 对文献中的词汇进行建模,而 PHITS 对文献的引文进行建模,在文献与引文之间引入主题空间,认为施引文献是被引文献具有文档中特定主题参数的多项式分布,运用 EM 算法对模型参数进行极大似然估计。该模型证实了引入引文后能够改进对文档的分类,可揭示一篇引文在特定主题条件下被引的可能性,也可计算出一篇引文的主题概率分布来识别主题特异性的引文,但是该模型未能抽取主题与词的分布。

随后,D. Cohn 和 T. Hofmann[8]提出了 PLSA-PHITS 联合主题模型,利用 PLSA 与 PHITS 基于同一因子分解,并共享同一个文档 - 主题的混合分布,由此引入了共同的潜在主题空间,能够同时抽取主题与词的分布、主题与引文的分布。该模型利用 EM 方法进行参数推断,产生的主题更加稳定,相比 PLSA 或 PHITS,分类效果更好。

之后国内外学者利用文本和引文数据在 LDA 基础上进行了扩展建模,如 E. Erosheva 等[9]提出了 Mixed-membership model,后期也有学者将该模型称为

Link-LDA 模型。该模型认为一篇文献是一个词袋的同时也是一个引文袋,引文与单词的生成相同,并共用同一个文档与主题分布,利用 LDA 过程分别产生主题与词的分布、主题与引文的分布。E. Erosheva 等利用该模型识别了 PNAS 生命科学领域 12 036 篇文献中的主题,与该数据集自身的学科分类相比,能够更细致地对这些文献进行分类。此外,有学者从施引文献的角度计算引用主题分布,如 T. Nguyen 等[10]建立了 CitationLDA＋＋模型,利用 LDA 获取主题与词分布,并作为先验知识用于模型的推断过程,在引用的主题分布计算中,从引文网络中获得施引文献集,对于每一篇文献,从先验知识中获得 top-k 主题,采用 Hellinger 距离计算施引文献主题与 top-k 主题间的相似度,来获得引用主题的分布。

随后,有学者在 Link-LDA 的基础上进行了扩展,如 Y. Li 等[11]提出了 cc-LDA 与 cp-LDA 模型, cc-LDA 模型与 Link-LDA 相近,但对于每一篇引文的处理过程不同,除抽取主题与引文的分布外,增加了抽取引文与词的分布;cp-LDA 模型引入了引文出现的位置,将一篇文章分成两部分,即"Introduction and Related Work"和"Others",由 beta 分布生成引文的位置。除此之外,一些学者尝试将引文与文献的其他元数据如作者进行联合建模,典型的有 Y. Tu 等[12]提出的引文作者主题模型 CAT,对单词、作者、引文因素联合建模;Z. Lu 等[13]提出 Collective Topic Model,在 PLSA 的基础上引入作者、论文发表地点、引文关系,利用共被引关系评价基于主题的论文影响力。

另外,还有一些学者抽取引文的主题分布,如 Z. Guo 等[14]构建了 CT 模型,先抽取文档与引文的分布,然后抽取引文与主题的分布,并且通过有向图的随机游走来捕捉间接引用关系。对该模型的性能评价采用了 Cora 数据库中的 9 998 篇文献,与其他模型进行主题聚类效果的对比,CT 模型要优于 PLSI、PHITS、LDA、PLSA-PHITS。X. Huang 等[15]设计了主题敏感的有影响力的论文识别模型(Model for Topic-sensitive Influential Paper Discovery, MTID),抽取施引文献的主题分布,并对论文在不同主题下的重要性进行建模。也有学者从分层的角度出发,如 H. Zhou 等[16]提出的 Citation-Content-LDA 模型,分为两层:第一层利用引文生成父主题,抽取的父主题代表了引文的聚类;第二层从第一层产生的每个父主题中抽取生成子主题。由于引用关系的数量比词的数量少,该模型可减小计算复杂度。

上述模型为参数化的模型,也有研究人员构建了非参数模型,如 K. W. Lim 和 W. Buntine[17－18]构建了引文网络主题模型 CNTM,引入作者、引文和文本内容,在泊松混合主题链接模型(Poisson Mixed-topic Link Model, PMTLM)[19]和作者主题模型(Author-Topic, AT)[20]模型的基础上扩展出非参数模型。H. Bai 等[21]提出了神经相关主题模型(Neural Relational Topic Model, NRTM),可同时利用主题和引文网络之间的潜在相关性。如表 1 所示:

表 1　基于主题与引文分布的主题模型

时间(年)	作者	模型	基础模型	参数估计与推断算法
2000	D. Cohn 等	PHITS	PLSA	EM 算法
2001	D. Cohn 等	PLSA＋PHITS	PLSA、PHITS	EM 算法
2004	E. Erosheva 等	Mixed-membership model	LDA	变分推理
2009	Z. Guo 等	CT	PLSA	EM 算法
2010	Y. Tu 等	CAT	LDA	Gibbs 采样
2014	K. W. Lim 等	CNTM	AT、PMTLM	MH 算法
2017	H. Zhou 等	Citation-Content-LDA	LDA	Gibbs 采样
2017	Y. Li 等	cc-LDA、cp-LDA	LDA	Gibbs 采样
2018	T. Nguyen 等	CitationLDA＋＋	LDA	Gibbs 采样

3.2　基于被引与施引主题间关系的主题模型

该类主题模型侧重于研究施引文献的主题与被引文献的主题之间的关系,通过选择是否从被引文献的主题分布中抽取主题,或者通过共用同一个主题－引文分布等多个角度揭示了被引文献的主题分布对施引文献主题分布的影响。该类模型包括 Copycat、Citation Influence Model(CIM)、Pairwise Link-LDA、Link-PLSA-LDA、Inheritance Topic Model(ITM)、Relational Topic Model(RTM)、cite-LDA、cite-PLSA-LDA、TERESA、Bernoulli Process Topic Model(BPT)、Bi-Citation-LDA、RefTM、Latent Topical Authority Indexing(LTAI)等。

L. Dietz 等[22]提出了 Copycat 与 CIM 模型,Copycat

模型中，施引文献中的每个主题从引文的主题混合中抽取，施引文献中的每一个词均与其引文进行关联，由此被引文献的主题分布影响着施引文献的主题，模型解释了被引和施引之间、共被引、耦合文献之间的依赖性。但由于该模型强制施引文献中的每个单词与一篇被引文献关联，两者在实际中并不能进行完全匹配，在被引文献的主题中会引入新的词，并且该模型不能揭示创新主题或正在发展中的主题。而 CIM 模型克服了这种限制，施引文献可选择是从引文的主题分布中抽取主题，或从其自身的主题分布中抽取，通过伯努利分布来进行选择，在实证研究中，CIM 模型的预测性能要优于 Copycat 模型。但该模型只能进行简单的双向图，未能处理复杂的引文网络。

M. Kim 等[23]在 CIM 模型基础上引入被引文献的 PageRank 值，来计算引用强度，通过该引用强度值来设定阈值，建立加权的引文网络进行主题扩散分析。随后，Z. Guo 等[24]提出的伯努利主题模型 BPT，认为同一篇文献扮演两个不同的角色，即文献本身和被引文献，作为被引文献时，主题的抽取与 LDA 相同，而对于文献本身的研究主题，其分布是引文的主题混合分布，引文网络的多层次结构通过随机的伯努利过程捕获。BPT 在困惑度上要优于 LDA、Link -LDA、Copycat 和 CIM。T. Masada 等[25]提出的 TERESA 模型与 BPT 模型类似。

也有学者将施引文献和被引文献构成一个文献对在主题模型中联合建模，典型的模型有 R. M. Nallapati 等[26]提出的 Pairwise -Link -LDA 与 Link -PLSA -LDA 模型，Pairwise-Link-LDA 模型结合了 LDA 和混合隶属度随机块模型（Mixed Membership Stochastic Block Models，MMSB）的优势，可以对任意链接结构进行建模。MMSB 最初用于蛋白质与蛋白质相互作用的建模，对于每一对蛋白质，分别抽取蛋白质的主题，两个蛋白质之间相互作用的有无由伯努利分布来生成。R. M. Nallapati 等将这种模型扩展至文本中，将文献看作蛋白质，对每一对文献的主题通过伯努利过程选择每对主题之间是否存在引用关系。MMSB 中，蛋白质相互作用是对称的，由于引用具有方向性，R. M. Nallapati 等采用文献的时间戳来为每一对文献分配方向性，在该模型中，单词的生成过程和 LDA 主题模型一致，施引文献与被引文献共用同一个主题分布。该模型虽能够清楚地揭示出施引文献和被引文献的主题关系，但由于它需要对每一对文献主题之间的引用关系进行计算，因此在当文档量较大时，计算成本较高，其扩展性受到限制。针对该问题，R. M. Nallapati 提出了 Link -PLSA-LDA 模型，结合了 PLSA 和 LDA 的优势。对于所有的被引文献，采用 PLSA 获取主题与引文的分布，对于每篇施引文献，利用 Link -LDA 获取主题与词的分布，基于 PLSA 的主题与引文分布抽取施引文献的主题与引文分布，不需要计算每一对文档，该模型保留了 Link -LDA 的可扩展性优势，在最大似然性和链接预测方面要优于 Pairwise -Link -LDA 模型，但 Pairwise -Link -LDA 模型在语义层面的揭示要优于 Link -PLSA -LDA 模型。

另外，还有学者提出 Pairwise -Link -LDA 模型中主题与词、主题与引用是分开产生的，不能保证识别的主题能同时能表征词与引用关系。针对该问题，J. Chang 和 D. M. Blei[27]进行了改进提出了相关主题模型（Relational Topic Model，RTM），施引文献与被引文献各自的生成过程与 LDA 相同，识别的施引文献与被引文献的主题之间通过伯努利过程选择是否存在关联关系。利用该模型对 Cora 2 708 篇数据进行主题识别，相比 LDA，RTM 对于施引和被引关系的识别精准性提高了 80%。随后，L. S. L. Tan 等[28]提出 LMV 模型（LDA MMSB Visibility），与 Pairwise-Link-LDA 相近，但对于每一篇引文，引入 beta 分布来产生与施引文献之间的关联关系，其预测性能优于 Pairwise -Link -LDA 和 RTM 模型。Q. He 等[29]提出 ITM 模型，对处于 t 时间的施引文献及其引文进行建模，来描绘主题之间的继承依赖性，分析主题随时间的演变，每篇文献中的词选择从引文的主题或施引文献的主题中生成。J. Shen 等[30]构建 RefTM 模型，与 ITM 的建模思想类似，在 RefTM 模型的基础上衍生出 J-Index 来评估文献在主题层面的学术影响力。L. Huang 等[31]提出 Bi -citation -LDA，被引文献采用 Link-LDA 模型同时生成主题与词、主题与引文的分布，若这篇文献被引用，对施引文献抽取主题与其引文的分布，该模型能够整合最新的文献，从而识别从被引文献流向施引文献的高影响力的主题。

也有一些模型在引入引文对关系的同时，加入了作者元数据进行建模，如 J. Kim 等[32]提出的 LTAI 模型，在每一对引文关系中引入作者的分布，且其中计算引文影响的参数服从狄利克雷分布。T. Dai 等[33]建立了作者链接社区的引文推荐主题模型（Topic Model with Author Link Community for Citation Recommendation），引入作者和引文信息，获取作者引文的分布、合作作者分布、施引文献与被引文献之间的关联关系，其引文推荐性能优于 Link-PLSA-LDA 和 RTM 模型。如表 2 所示：

表2　基于被引与施引主题间关系的主题模型

时间(年)	作者	模型	基础模型	参数估计与推断算法
2007	L. Dietz 等	Latent Dirichlet Allocation model、Copycat Model、CIM	LDA	Gibbs 采样
2008	R. M. Nallapati 等	Pairwise Link-LDA、Link-PLSA-LDA	LDA、PLSA、MMSB	变分推理
2009	Q. He 等	ITM	LDA	Gibbs 采样
2010	J. Chang 等	RTM	LDA	变分推理
2012	T. Masada 等	TERESA	LDA	变分推理
2014	Z. Guo 等	BPT	LDA	变分推理
2015	L. S. L. Tan 等	LMV	LDA、MMSB	变分推理
2016	L. Huang 等	Bi-Citation-LDA	LDA	Gibbs 采样
2016	J. Shen 等	RefTM	LDA	Gibbs 采样
2017	J. Kim 等	LTAI	LDA	变分推理、EM 算法

3.3　基于引用内容的引用主题模型

H. Small[34]在 1982 年提出引用内容(citation context)的定义,指出现在参考文献标签周围的文本内容。国内外研究人员利用引用内容开展了主题提取、主题聚类等方面的探索性应用研究。B. Aljaber 等[35]发现引用内容的主题词可以较好地识别研究主题用于文献的聚类。L. Bornmann 等[36]发现引用内容比题目和摘要中提取的关键词在语义上更接近于学者文章中的研究内容。M. Doslu 等[37]利用引用内容构建有向的主题词标引的引文网络,利用 HITS 算法对特定主题的论文进行排序,识别基于主题的重要文献;S. Liu 等[38]利用 LDA 识别引用内容主题,发现引用内容主题比引文自身主题涉及范围更广。杨春艳等[39]利用 Labeled-LDA 结合的主题模型抽取引用内容主题,发现引用内容可以消除全文存在的“噪音”,并能覆盖尽可能多的主题内容。X. Liu 等[40]采用 Labeled-LDA,构建了被引文献和施引文献间基于引用内容的网络图,解决了引用原因以及引文贡献值的问题。

现有的研究表明引用内容相对于引文分析,包含了更丰富的主题相关的语义信息。相比引文来说,研究对象不再以文献为最小单位,而是细化到文献中的知识元,将节点属性和节点间的关系赋予新的理解,而将引用内容引入到主题模型中进行建模的研究相对较少。

S. Kataria 等[41]在 Link-LDA 和 Link-PLSA-LDA 基础上引入了引用内容,提出 cite-LDA、cite-PLSA-LDA 模型。其模型假设引用内容中,词和被引文献的选择是相互独立的,即对于引用内容中的词,同时抽取主题与词、主题与引文的分布。Cite-PLSA-LDA 模型中,施引文献采用了 Cite-LDA 的生成过程,而被引文献则是利用 PLSA 抽取被引文献的主题与引文的分布。在模型的性能评估实验中,利用这两种模型分别对 CiteSeer 3 312 篇文献数据进行主题识别,Cite-LDA 与 Link-LDA、Link-PLSA-LDA 效果近似,而 Cite-PLSA-LDA 要优于其他三个模型。

4　结论与展望

本文系统地梳理了近年来国内外学者们提出的利用引文构建的主题模型的发展现状,详细分析和对比了各模型的思想及生成过程,为情报分析中主题识别和演化分析的方法选择提供参考,也为基于上述模型的进一步改进和完善提供思路。

从上述利用引文构建的主题模型的研究中可以看出,近几年来随着学者们的不断研究与探索,对利用引文构建的主题模型的研究工作主要集中在模型的扩展、改进和优化方面。研究角度包括研究主题与引文分布,侧重研究被引与施引主题间关系。参与主题识别的词汇的来源包括了施引文献和被引文献,随着全文本分析的发展,也出现了引入引用内容的主题模型。现有的模型均表明,引入引文信息后,可改进对主题的识别,能同时准确地抽取主题的关键词分布和关键文献分布,能够获得更完整的主题内容,改进对文档的分类,可以关联施引文献和被引文献之间主题间的关系及影响,可以为主题演化分析提供重要的量化分析作用。

然而,现有的利用引文构建的主题模型的研究仍然存在一些问题,从上文的对比分析中,可看出基于主题与引文分布的主题模型中,由于只对文档引用特征进行建模,未能对施引文献与被引文献文本之间的主题关系进行建模,不能展示被引文献与施引文献的主题继承性,底层的生成过程相对简单而不能解释语料库中引文结构和各种现象。基于被引与施引主题间关

系的主题模型能够提示被引文献与施引文献的主题层面的关联关系,然而从主题识别采用的词汇来源看,目前多集中在从被引文献的标题、摘要等来抽取词汇,而采用引用内容来表征被引文献主题的研究并不充分。可见,利用引文构建主题模型仍需要进一步推动,未来的发展趋势可能会向以下方向延伸。

4.1 引入引用内容的主题模型研究及扩展

目前对引用内容的主题模型研究中,引入到主题模型中进行建模的研究较少,主要是应用已成熟的模型识别引用内容的主题。随着引用内容分析和自然语言处理两个研究领域交叉的深入,使用引用内容对引文主题进行语义分析和自动分析将会更加深入,这将会加强引文分析的深度,特别是对语义理解的程度。此外,随着互联网技术的进一步发展和开放获取运动的兴起,全文数据成为了易获取、易解析的数据来源,同时包含了更加丰富的文本信息。如 xml 格式的全文数据利用结构化标记语言,对全文中引用内容等信息进行了标记,这些均会进一步促进引用内容分析的发展。如何对这些全文数据提取引用内容文本元素,进行文本挖掘和语义分析,建立合适的主题模型来提取和识别潜在的主题结构和演化信息,应该进行深入的研究。

4.2 利用引文构建的主题模型的性能优化和评价方法

利用引文构建的主题模型在性能优化方面需要更高效的算法。目前大部分模型是将词项或引文空间变换到主题空间,已有的模型多集中在 PLSA 和 LDA 的扩展方面,对参数的估计和推断算法上多采用 EM 算法、变分推理、Gibbs 采样等方法。此外,该类模型中,特别是基于被引与施引主题间关系的主题模型,会引入新的潜在变量,模型的运行时间通常会增加。如 Link-LDA 和 Link-PLSA-LDA 中,Gibbs 单次采样时间的复杂度与语料库中链接的数量呈线性相关,在链接过大时该模型将会受限。如何设计针对该类模型的性能优化方法,以及如何在降低复杂度和保证主题词效果之间寻求平衡需要进行深入的研究。此外,目前对利用引文构建的主题模型的评价采用复杂度比较、召回率方法评估模型的效果,个别模型采用了 AUC、精确率、主题一致性(topic coherence)、F1 Score 的方法,对模型效果的评估方法还可从多个角度进行扩充。

4.3 利用引文构建的主题模型的应用研究

利用引文构建的主题模型目前主要应用在主题识别、主题演化、文本聚类、链接预测、引文推荐等方面。结合文本的主题信息与引用关系进行建模,在发现高质量的主题的同时,还可预测引文的强度,发现主题内部之间的继承与演化关系。这种主题间的关联关系可丰富主题影响力评价,或结合已有的计量学指标如 h 指数和影响因子,扩展研究基于主题关联关系的学术影响力评价指标。该类模型本质上是一种对具有链接信息的文本概率建模的方法,可以应用在文本挖掘的多个方面。已有的研究中,应用的对象除科技文献外,个别学者也将其应用至网页信息以及 blog 数据中,表明了该类模型可扩展应用到带有链接的多种语料中,但对应用的效果评价还需要进行更深入的研究。

参考文献:

[1] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3(Jan):993 - 1022.

[2] 张金松. 基于引文上下文分析的文献检索技术研究[D]. 大连:大连海事大学, 2013.

[3] HOFMANN T. Probabilistic latent semantic analysis[C]//Association for Uncertainty in Artificial Intelligence. Fifteenth conference on uncertainty in artificial intelligence. Stockholm: Morgan Kaufmann, 1999:289 - 296.

[4] 范云满, 马建霞. 利用 LDA 的领域新兴主题探测技术综述[J]. 现代图书情报技术, 2012, 28(12):58 - 65.

[5] KAWAMAE N. Trend analysis model: trend consists of temporal words, topics, and timestamps[C]//International conference on web search and data mining. Heng Kong: Association for Computing Machinery, 2011:317 - 326.

[6] ROSEN-ZVI M, GRIFFITHS T, STEYVERS M, et al. The author-topic model for authors and documents[C]//Association for Uncertainty in Artificial Intelligence. Proceedings of the 20th conference on uncertainty in artificial intelligence. Banff: Association for Uncertainty in Artificial Intelligence Press, 2012:487 - 494.

[7] COHN D, CHANG H. Learning to probabilistically identify authoritative documents[C]//Association for Computing Machinery. Proceedings of the seventeenth international conference on machine learning. San Francisco: Morgan Kaufmann Publishers, 2000:167 - 174.

[8] COHN D, HOFMANN T. The missing link: a probabilistic model of document content and hypertext connectivity[C]//Neural Information Processing Systems Foundation. Advances in neural information processing systems 13. Cambridge: NIPS, 2000:430 - 436.

[9] EROSHEVA E, FIENBERG S, LAFFERTY J. Mixed-membership models of scientific publications [J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101 (1):5220 - 5227.

[10] NGUYEN T, DO P. CitationLDA plus: an extension of LDA for

discovering topics in document network[C]//Association for Computing Machinery. International symposium on information and communication technology. Danang City: Association for Computing Machinery, 2018:31－37.

[11] LI Y, HE J, LIU H. Topic analysis and influential paper discovery on scientific publications[C]//14th web information systems and applications conference. Liuzhou: IEEE, 2017:68－73.

[12] TU Y, JOHRI N, ROTH D, et al. Citation author topic model in expert search[C]// Association for Computational Linguistics. International conference on computational linguistics: posters. Beijing: Association for Computational Linguistics, 2010: 1265－1273.

[13] LU Z, MAMOULIS N, CHEUNG D. A collective topic model for milestone paper discovery[C]// Association for Computing Machinery. Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval. Queensland: Association for Computing Machinery, 2014:1019－1022.

[14] GUO Z, ZHU S, CHI Y, et al. A latent topic model for linked documents[C]// Association for Computing Machinery. Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. Boston: Association for Computing Machinery, 2009:720－721.

[15] HUANG X, CHEN C, PENG C, et al. Topic-sensitive influential paper discovery in citation network [C]//PacificAsia conference on knowledge discovery & data mining. Melbourne: Springer, 2018:16－28.

[16] ZHOU H, HUIMIN Y, ROLAND H. Topic discovery and evolution in scientific literature based on content and citations [J]. Frontiers of information technology & electronic engineering, 2017, 18(10):1511－1532.

[17] LIM K W, BUNTINE W. Bibliographic analysis on research publications using authors, categorical labels and the citation network [J]. Machine learning, 2016, 103(2):185－213.

[18] LIM K W, BUNTINE W. Bibliographic analysis with the citation network topic model[C]//Asian conference on machine learning. JMLR Workshop and conference proceedings. Nha Trang City: Springer, 2014, 39:142－158.

[19] ZHU Y, YAN X, GETOOR L, et al. Scalable text and link analysis with mixed-topic link models [C]//Association for Computing Machinery. Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. Chicago: Association for Computing Machinery, 2013, 47:473－481.

[20] YAN L, NICULESCU-MIZIL A, GRYC W. Topic-link LDA: joint models of topic and author community[C]//Association for Computing Machinery. Proceedings of the 26th annual international conference on machine learning. Montreal: Association for Computing Machinery, 2009:665－672.

[21] BAI H, CHEN Z, LYU M. Neural relational topic models for scientific article analysis [C]// Association for Computing Machinery. Proceedings of the 27th ACM international conference on information and knowledge management. Torino: Association for Computing Machinery, 2018:27－36.

[22] DIETZ L, BICKEL S, SCHEFFER T. Unsupervised prediction of citation influences[C]// Association for Computing Machinery. Proceedings of the 24th international conference on Machine learning. Corvalis: Association for Computing Machinery, 2007:233－240.

[23] KIM M, BAEK I, SONG M. Topic diffusion analysis of a weighted citation network in biomedical literature [J]. Journal of the Association for Information Science and Technology, 2018, 69(2):329－342.

[24] GUO Z, ZHANG Z M, ZHU S, et al. A two-level topic model towards knowledge discovery from citation networks [J]. IEEE transactions on knowledge & data engineering, 2014, 26(4):780－794.

[25] MASADA T, TAKASU A. Extraction of topic evolutions from references in scientific articles and its GPU acceleration[C]//Association for Computing Machinery. International conference on information and knowledge management. Maui: Association for Computing Machinery, 2012:1522－1526.

[26] NALLAPATI R M, AHMED A, XING E P, et al. Joint latent topic models for text and citations[C]//Association for Computing Machinery. Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. Las Vegas: Association for Computing Machinery, 2008:542－550.

[27] CHANG J, BLEI D M. Hierarchical relational models for document networks [J]. Annals of applied statistics, 2010, 4(1):124－150.

[28] TAN L S L, HUI C A, TIAN Z. Topic-adjusted visibility metric for scientific articles [J]. The annals of applied statistics, 2016, 10(1):1－31.

[29] HE Q, CHEN B, PEI J, et al. Detecting topic evolution in scientific literature: how can citations help? [C]//Association for Computing Machinery. Proceedings of the 18th ACM conference on information and knowledge management. Hong Kong: Association for Computing Machinery, 2009:957－966.

[30] SHEN J, SONG Z, LI S, et al. Modeling topic-level academic influence in scientific literatures[C]//Association for the Advancement of Artificial Intelligence. The workshops of the thirtieth AAAI conference on artificial Intelligence. Phoenix: Association for the Advancement of Artificial Intelligence, 2016:1－7.

[31] HUANG L, LIU H, HE J, et al. Finding latest influential research papers through modeling two views of citation links[C]//Asia-pacific web conference, Web technologies and applications. Suzhou: Springer, 2016:555－566.

[32] KIM J, KIM D, OH A. Joint modeling of topics, citations, and topical authority in academic corpora [J]. Transactions of the as-

sociation for computational linguistics, 2017, 5(1):191-204.
[33] DAI T, ZHU L, CAI X, et al. Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network[J]. Journal of ambient intelligence and humanized computing, 2018, 9(5):957-975.
[34] SMALL H. Citation context analysis [J]. Progress in communication sciences, 1982, 3(9):287-310.
[35] ALJABER B, STOKES N, BAILEY J, et al. Document clustering of scientific texts using citation contexts [J]. Information retrieval, 2010, 13(2):101-131.
[36] BORNMANN L, HAUNSCHILD R, HUG S E. Visualizing the context of citations referencing papers published by Eugene Garfield: a new type of keyword co-occurrence analysis[J]. Scientometrics, 2018, 114(2):427-437.
[37] DOSLU M, BIGNOL H O. Context sensitive article ranking with citation context analysis [J]. Scientometrics, 2016, 108(2):653-671.
[38] LIU S, CHEN C. The differences between latent topics in abstracts and citation contexts of citing papers [J]. Journal of the American Society for Information Science and Technology, 2013, 64(3):627-639.
[39] 杨春艳, 潘有能, 赵莉. 基于语义和引用加权的文献主题提取研究[J]. 图书情报工作, 2016, 60(9):131-138.
[40] LIU X, ZHANG J, GUO C. Full-text citation analysis: a new method to enhance scholarly networks [J]. Journal of the American Society for Information Science and Technology banner, 2013, 64(9):1852-1863.
[41] KATARIA S, MITRA P, BHATIA S. Utilizing context in generative bayesian models for linked corpus[C]//Association for Computing Machinery. Twenty-fourth AAAI conference on artificial intelligence. Atlanta: Association for Computing Machinery, 2010: 1340-1345.

作者贡献说明:
邹丽雪:设计研究思路,撰写及修改论文;
王丽:修订论文;
刘细文:指导研究思路,修订论文。

Research Advances of Citation Based Topic Models

Zou Lixue[1,2] Wang Li[1,2] Liu Xiwen[1,2]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190
[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] A wide variety of topic models has been developed with improved algorithm. This paper aims to study the research advances, generation process and algorithm of citation based topic models. Additionally, we discuss the application in the text of academic articles and research areas in the future. [**Method/process**] Based on the data of Web of Science and CNKI database, we collected articles of citation based topic models. In these articles, we selected several representative articles after manual interpretation to analyze the generative process, parameter estimation and inference methods in these citation based topic models. [**Result/conclusion**] Currently, there are mainly three types of citation based topic models. This includes the topic models which focus on the topic-citation distribution, while other topic models mainly study the relationship between the citing documents and the cited documents. Besides, citation context based topic models are also available. Additionally, more complete topic content can be detected after introducing citation information into the topic models. Moreover, most of the models are the variants of LDA and PLSA. In future, incorporating citation context information into topic models, improving the inference methods and applying the models are some of the future directions.

Keywords: topic model citation topic detection citation context